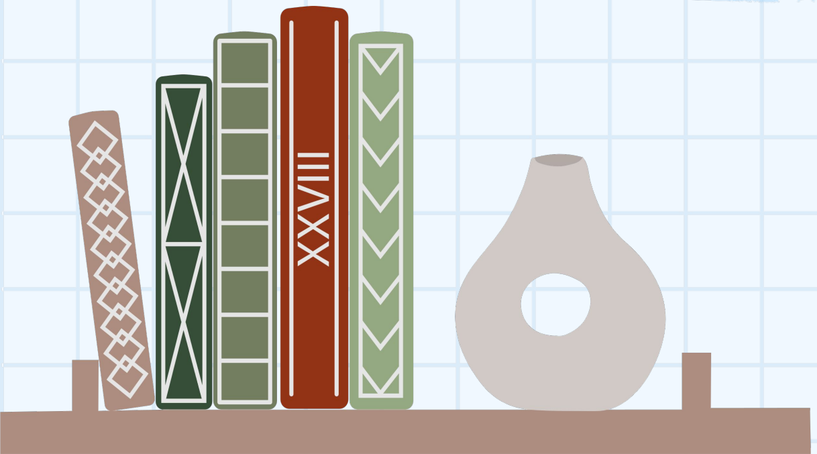
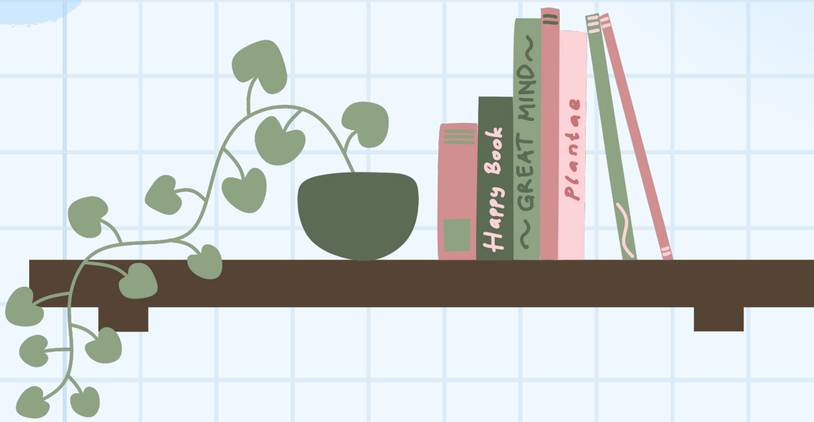
Powered by
Futurense

BS./BSC.IN

Applied ai and Data Science

Basics of Data Analytics





Let's dive into and learn:

1 Data Transformations

Powered by Futurense

Data Transformations

Powered by

Futurense

- After cleaning the data, we often have to make some more transformation to the data before starting the analysis.
- Transformations look like
    - Changing the scale of variables
    - Encoding categorical variables

Transforming Continuous

- Changing the scale of the variables
    - Make them more interpretable or comparable
- Managing the distribution of the variables
    - Handling skewness of variables

Changing the scale of the

Powered by Futurense

Standardization (Z-score Normalization)

$$z_x = \frac{x_i - \bar{x}}{s}$$

Here

x_i is the original value

$\bar{x}$ is the sample mean

s is the sample standard deviation

Changing the scale of the

- Used when variables have different units of measurement and need to be on a common scale
- When a few large numbers dominate the dataset.

Changing the scale of the

- Min-max Scaling (Normalization)

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Here,

x_i is the original value

x_{min} is the minimum value in the sample

x_{max} is the maximum value in the sample

Changing the scale of the

- Rescales data to a fixed range, typically [0,1]
- Used when working with fixed-interval data (e.g. percentages and proportions)
- When preserving relative relationships between values is important
- Caution! - This method is sensitive to outliers, as extreme values determine the range

Changing the scale of the

- Logarithmic Transformations
  - $x_i' = \log(x_i)$ when all the values are positive
  - $x_i' = \log(1 + x_i)$ when the data contains zeros

Powered by Futurense

Changing the scale of the

- Logarithmic Transformations are useful for handling highly skewed data.
- They compress large values while expanding small values

Encoding Categorical Variables

Powered by

Futurense

- There are multiple ways to transform categorical variables into numerical representations
- Essential for handling data and conducting the analysis correctly

Encoding Categorical Variables

Powered by Futurense

One-Hot Encoding (Dummy Variables)

- Creates binary indicator variables for each category
- For a categorical variable with *k* unique values, create *k* binary columns

$$x_i = \begin{cases} 1 \; if \; x_i = k \\ 0 \; otherwise \end{cases}$$

Encoding Categorical Variables

Powered by Futurense

One-Hot Encoding (Dummy Variables)

- Used to encode variables with small number of categories
- E.g. Gender (male/female) , Place of stay (Rural/Urban)

Encoding Categorical Variables

Powered by Futurense

Label Encoding

- Assigns integer values to categories
- Category → Integer

Category	Encoded Value
Red	1
Blue	2
Green	3

Encoding Categorical Variables

Powered by Futurense

Label Encoding

- Used when you have discrete variables without any logical ordering

Category	Encoded Value
Red	1
Blue	2
Green	3

Encoding Categorical Variables

Powered by Futurense

Ordinal Encoding

- Maps categories to ordered integer values based on logical ranking

Category	Encoded Value
Low	1
Medium	2
High	3

Encoding Categorical Variables

Powered by Futurense

Ordinal Encoding

- Used when you have ordinal variables with meaningful order

  (e.g., satisfaction ratings).

Category	Encoded Value
Low	1
Medium	2
High	3

Recap

Powered by

Futurense

Summarized version of lecture

Powered by Futurense

Thank you